Fast Heterogeneous Federated Learning with Hybrid Client Selection (Supplementary Material)

Duanxiao Song[*1] **Guangyuan Shen**[*1,2] **Dehong Gao**[*1,2] **libin yang**[†1] **Xukai Zhou**[1] **Shirui Pan**[3] **Wei Lou**[4]
Fang Zhou[1]

[1]Department of Cybersecurity, Northwestern Polytechnical University, China
[2]Alibaba Group, China
[3]School of Information and Communication Technology, Griffith University, Australia
[4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

1 EXTRA ILLUSTRATION

1.1 DIRICHLET DISTRIBUTION ILLUSTRATION

Given different values of the concentration parameter α for the Dirichlet distribution, datasets with different degrees of heterogeneity can be generated. In particular, higher values of α lead to a more uniform distribution, indicating that each client has an almost equally weighted combination of labels. Lower values of α imply weights concentrated more heavily on only one of the labels, or more extreme label membership. Table 1 is an example of Dirichlet distribution used in the experiments. As shown in the Table 1, the data distribution on each client is different, and in the case of extreme non-IID, i.e., $\alpha \to 0$, most of the data are concentrated under only one label, while the amount of others is almost zero.

1.2 GRADIENT COMPRESS ALGORITHM

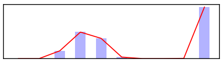
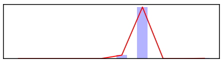
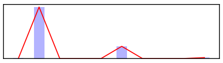
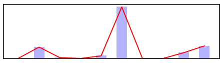
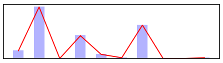
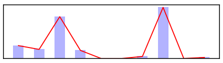
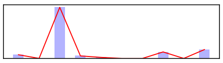
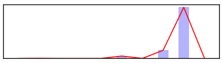
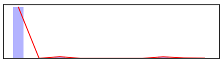
Algorithm 1: GC(Gradient Compress)

```
Input: Raw updates in the $t^{th}$ round of the $k^{th}$ client $G_t^k = \{g_1, g_2, \dots, g_d\}$
Initialize Randomly select $d'$ $g_i$ as the group centers $\{x_1, x_2, \dots, x_{d'}\}$;
Initialize $C_j = \varnothing$ $(1 \le j \le d')$;
repeat
    for each $g_i,\ i = 1, 2, \dots, d$ do
        $\lambda_i = \arg\min_{j \in \{1,2,\dots,d'\}} \|g_i - x_j\|_2$;
        $C_{\lambda_i} = C_{\lambda_i} \bigcup \{g_i\}$;
    end
    for each cluster $j = 1, 2, \dots, d'$ do
        Calculate new center $x'_j = \frac{1}{|C_j|} \sum_{g_i \in C_j} g_i$;
        $x_j \leftarrow x'_j$;
    end
until $\forall\, j = \{1, 2, \dots, d'\}, x'_j = x_j$;
Output: $\boldsymbol{X}_t^k = \{x_1, x_2, \cdots, x_{d'}\}$;
```

[*] The first three authors contribute equally to this work.
[†] Contact author.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 1: The actual Dirichlet Distribution (non-IID) generated from CIFAR-10 with $\alpha = 0.001$

Client ID	**Numbers of Samples in the Classes**										**Distribution**
	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	
k=0	2	1	33	117	100	6	1	**0**	1	239	
k=1	**0**	**0**	**0**	1	1	29	**467**	**0**	**0**	2	
k=2	2	**397**	5	1	2	86	**0**	1	**0**	6	
k=3	1	**0**	**0**	**0**	**0**	3	**0**	**0**	125	**371**	
k=4	1	67	5	**0**	15	**304**	**0**	**0**	34	74	
...	...	...	...	...	...	...	...	...	...	...	...
k=95	32	213	**0**	94	17	3	138	**0**	**0**	3	
k=96	51	36	166	32	**0**	**0**	8	203	**0**	4	
k=97	25	**0**	**347**	17	7	**0**	**0**	44	**0**	60	
k=98	**0**	1	**0**	2	1	18	3	60	**413**	2	
k=99	**465**	**0**	4	2	2	3	4	14	4	2	

1.3 KEY LEMMAS

Following Li et al. [2019], we present necessary assumptions and extra notations that we used to prove the convergence of FedAvg with random client selection.

Assumptions. The convergence of FedAvg with random sampling scheme has been derived in Li et al. [2019]. The proof relies on the assumptions as follows. Assumptions 3 and 4 have been given by Zhang et al. [2013], Stich [2018], Stich et al. [2018][1].

Assumption 1 (L-Smooth)**.** *$\forall$ $\mathbf{v}$ and $\mathbf{w}$, $k = 1, \cdots, N F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$ where the $\mathbf{v}$,$\mathbf{w}$ are different model parameters.*

Assumption 2 (Strongly Convex)**.** *$\forall$ $\mathbf{v}$ and $\mathbf{w}$, $k = 1, \cdots, N F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$ where the $\mathbf{v}$,$\mathbf{w}$ are different model parameters.*

Assumption 3 (Bounded Variance)**.** *Let ξ_t^k be sampled from the k^{th} device's local data uniformly at random. The variance of stochastic gradients in each device is bounded:*

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right) - \nabla F_k\left(\mathbf{w}_t^k\right)\right\|^2 \leq \sigma_k^2, \; \forall\, k = 1, \cdots, N$$

Assumption 4 (Bounded Expectation)**.** *The expectation of stochastic gradients in squared norm is bounded by G^2, i.e.,*

$$\mathbb{E}\left\|\nabla F_k\left(\mathbf{w}_t^k, \xi_t^k\right)\right\|^2 \leq G^2, \forall\, k = 1, \cdots, N, t = 1, \cdots T - 1$$

Additional Notation. We assume that FedAvg always activates all devices at the beginning of each round and then uses the parameters maintained in only a few sampled devices to produce the next-round parameter. This updating scheme is

[1] Note that the strong assumptions are only used for convergence analysis, and variance comparison does not require these assumptions.

equivalent to the original. Let $\mathcal{I}_E$ be the set of global synchronization, i.e., $\mathcal{I}_E = \{nE \mid n = 1, 2, \cdots\}$. If $t+1 \in \mathcal{I}_E$, i.e., the time step to communicate. Then the update of FedAvg with partial devices active can be described as: for all $k \in [N]$,

$$\mathbf{v}_{t+1}^k = \mathbf{w}_t^k - \eta_t \nabla F_k \left(\mathbf{w}_t^k, \xi_t^k\right) \tag{1}$$

$$\mathbf{w}_{t+1}^k = \begin{cases} \mathbf{v}_{t+1}^k & \text{,if } t+1 \notin \mathcal{I}_E \\ \text{average } \left\{\mathbf{v}_{t+1}^k\right\}_{k \in \mathcal{S}_{t+1}} & \text{,if } t+1 \in \mathcal{I}_E, \end{cases} \tag{2}$$

where $\mathcal{S}_{t+1}$ denotes the subset of $(t+1)^{th}$ round. Here, an additional variable $\mathbf{v}_{t+1}^k$ is introduced to represent the immediate result of one step SGD update from $\mathbf{w}_t^k$. We interpret $\mathbf{w}_{t+1}^k$ as the parameter obtained after communication steps. Let F^* and F_k^* be the minimum values of F and F_k, respectively. We use the term $\Gamma = F^* - \sum_{k=1}^N p_k F_k^*$ for quantifying the degree of non-IID, where the p_k denotes the aggregation weight. If the data are IID, then Γ goes to zero as the number of samples grows. If the data are non-IID, then Γ is nonzero, and its magnitude reflects the heterogeneity of the data distribution.

Lemma 1 (Results of one step SGD)**.** *Assume Assumption 1 and 2. If $\eta_t \leq \frac{1}{4L}$, we have*

$$\mathbb{E}\left\|\overline{\mathbf{v}}_{t+1} - \mathbf{w}^\star\right\|^2 \leq (1 - \eta_t \mu)\,\mathbb{E}\left\|\overline{\mathbf{w}}_t - \mathbf{w}^\star\right\|^2$$
$$+ \eta_t^2 \mathbb{E}\left\|G_t - \overline{G}_t\right\|^2 + 6L\eta_t^2\Gamma + 2\mathbb{E}\sum_{k=1}^N p_k \left\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\right\|^2$$

where $\Gamma = F^* - \sum_{k=1}^N p_k F_k^\star \geq 0$,

Lemma 2 (Bounding the variance)**.** *Assume Assumption 3 holds, and σ_k defined there. It follows that*

$$\mathbb{E}\left\|G_t - \overline{G}_t\right\|^2 \leq \sum_{k=1}^N p_k^2 \sigma_k^2,$$

where the G_t is the gradient vector of t^{th} round.

Lemma 3 (Bounding the divergence of $\left\{\mathbf{w}_t^k\right\}$)**.** *Assume Assumption 4, that η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. It follows that*

$$\mathbb{E}\left[\sum_{k=1}^N p_k \left\|\overline{\mathbf{w}}_t - \mathbf{w}_t^k\right\|^2\right] \leq 4\eta_t^2 (E-1)^2 G^2.$$

Lemma 4 (Unbiased sampling scheme)**.** *If $(t+1)^{th}$ round is the communication round, for our selection with $\mathcal{S}_t = \{i_1, \cdots, i_m\} \subset [N]$ we have*

$$\mathbb{E}\left[\mathbf{w}(\mathcal{S}_t)\right] = \mathbf{w}(\mathcal{K}),$$

where $\mathcal{K}$ denotes the population of clients.

Proof.

$$\mathbb{E}\left[\mathbf{w}(\mathcal{S}_t)\right] = \mathbb{E}_{\mathcal{S}_t}\sum_{k=1}^m \mathbf{w}_{i_k} = m\mathbb{E}_{\mathcal{S}_t}\left[\mathbf{w}_{i_1}\right] = m\sum_{k=1}^N p_k \mathbf{w}_k \tag{3}$$

□

Lemma 5 (Bounding the variance of $\mathbf{w}(\mathcal{S}_t)$)**.** *For $t+1 \in \mathcal{I}_E$, assume that η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. We have the following result assuming $p_1 = p_2 = \cdots = p_{m_h} = \frac{1}{N_h}$, the expected difference between $\overline{\mathbf{v}}_{t+1}$ and $\overline{\mathbf{w}}_{t+1}$ is bounded by*

$$\mathbb{E}_{\mathcal{S}_t}\left\|\overline{\mathbf{v}}_{t+1} - \overline{\mathbf{w}}_{t+1}\right\|^2 \leq \frac{4}{K}\eta_t^2 E^2 G^2$$

2 PROOF OF THEOREM

2.1 PROOF OF THEOREM 1 VARIANCE REDUCTION

Additional Notation. Divide the population $\mathcal{K}$ consisting of N clients into H clusters via clustering.

- N_h denotes the number of clients in h^{th} cluster, $s.t. \sum_{h=1}^{\mathrm{H}} N_h = N$
- m_h denotes the number of sampled clients from the h^{th} cluster
- m denotes the sample size, $s.t. \sum_{h=1}^{\mathrm{H}} m_h = m$
- $\mathbf{w}_{h_i}$ denotes the model update $\mathbf{w}$ of the i^{th} client in the h^{th} cluster
- $\mathbf{w}_h = \sum_{i=1}^{m_h} \frac{\mathbf{w}_{h_i}}{m_h}$ is the sampled averaged model update of the h^{th} cluster
- $\overline{\mathbf{w}} = \sum_{h=1}^{\mathrm{H}} \frac{m_h \mathbf{w}_h}{m}$ is the overall sampled averaged model update
- $\mathbf{W}_h = \sum_{i=1}^{N_h} \frac{\mathbf{w}_{h_i}}{N_h}$ is the averaged model update of the h^{th} cluster
- $\mathbf{W}(\mathcal{K}) = \sum_{h=1}^{\mathrm{H}} \sum_{i=1}^{N_h} \frac{\mathbf{w}_{h_i}}{N}$ is the averaged model update of entire set $\mathcal{K}$
- $\mathbf{w}_{cluster} = \frac{1}{N} \sum_{h=1}^{\mathrm{H}} N_h \mathbf{w}_h$ is an unbiased estimator of $\mathbf{W}(\mathcal{K})$
- $S^2 = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{w}_i - \mathbf{W}(\mathcal{K})\|_2^2 := \frac{1}{N} \sum_{i=1}^{N} \sigma^2$
- ${S_h}^2 = \sum_{i=1}^{N_h} \frac{\|\mathbf{w}_{h_i} - \mathbf{W}_h\|_2^2}{N_h - 1}$
- ${s_h}^2 = \sum_{i=1}^{m_h} \frac{\|\mathbf{w}_{h_i} - \mathbf{w}_h\|_2^2}{m_h - 1}$
- $Q_h = \frac{N_h}{N}$ is the proportion of clients in the h^{th} cluster
- $q_h = \frac{m_h}{m}$ is the proportion of sampled clients in the h^{th} cluster

Proof of Theorem 1. **Derive the Variance of Random Selection.** Assuming that each observation has variance σ^2, then we get

$$\mathbb{V}(\mathbf{w}_{rand}) = \mathbb{E}\left\|\overline{\mathbf{w}} - \mathbf{W}(\mathcal{K})\right\|_2^2 \tag{4}$$

$$= \frac{1}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}[\mathbf{w}_i - \mathbf{W}(\mathcal{K})]\right\|_2^2 \tag{5}$$

$$= \underbrace{\frac{1}{m^2}\mathbb{E}\left[\sum_{i=1}^{m}\|\mathbf{w}_i - \mathbf{W}(\mathcal{K})\|_2^2\right]}_{Quadratic\ Term} \tag{6}$$

$$+ \underbrace{\frac{1}{m^2}\mathbb{E}\left[\sum_{i}^{m}\sum_{\neq j}^{m}[\mathbf{w}_i - \mathbf{W}(\mathcal{K})]^T[\mathbf{w}_j - \mathbf{W}(\mathcal{K})]\right]}_{Cross-Product\ Term} \tag{7}$$

$$= \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{w}_i - \mathbf{W}(\mathcal{K})\right\|_2^2 \tag{8}$$

$$+ \frac{1}{m^2}\underbrace{\sum_{i}^{m}\sum_{\neq j}^{m}\mathbb{E}\left[[\mathbf{w}_i - \mathbf{W}(\mathcal{K})]^T[\mathbf{w}_j - \mathbf{W}(\mathcal{K})]\right]}_{Setting\ to\ K} \tag{9}$$

$$= \frac{1}{m^2}\sum_{i=1}^{m}\sigma^2 + \frac{K}{m^2} \tag{10}$$

$$= \frac{N-1}{Nm}S^2 + \frac{K}{m^2}, \tag{11}$$

where we set $K = \sum_i^m \sum_{\neq j}^m \mathbb{E}\left[[\mathbf{w}_i - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_j - \mathbf{W}(\mathcal{K})]\right]$ for convenience.

Find the Expression of K. In order to find K, we consider,

$$\begin{aligned}&\mathbb{E}\left[[\mathbf{w}_i - \mathbf{W}(\mathcal{K})][\mathbf{w}_j - \mathbf{W}(\mathcal{K})]\right] \\ &= \frac{1}{N(N-1)} \sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [(\mathbf{w}_l - \mathbf{W}(\mathcal{K})]\right].\end{aligned} \tag{12}$$

Meanwhile, we have,

$$\sum_{k=1}^N [\mathbf{w}_k - \mathbf{W}(\mathcal{K})] = \sum_{k=1}^N \mathbf{w}_k - N\mathbf{W}(\mathcal{K}) \tag{13}$$

$$= N\mathbf{W}(\mathcal{K}) - N\mathbf{W}(\mathcal{K}) = 0, \tag{14}$$

i.e.,

$$\left\| \sum_{k=1}^N [\mathbf{w}_k - \mathbf{W}(\mathcal{K})] \right\|_2^2 = 0. \tag{15}$$

And the left can be constructed as,

$$\begin{aligned}\left\| \sum_{k=1}^N [\mathbf{w}_k - \mathbf{W}(\mathcal{K})] \right\|_2^2 &= \sum_{k=1}^N \|\mathbf{w}_k - \mathbf{W}(\mathcal{K})\|_2^2 \\ &+ \sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_\ell - \mathbf{W}(\mathcal{K})]\right].\end{aligned} \tag{16}$$

Simplify it, we will get

$$0 = (N-1)S^2 + \sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_\ell - \mathbf{W}(\mathcal{K})]\right], \tag{17}$$

equal to,

$$\sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_\ell - \mathbf{W}(\mathcal{K})]\right] = -(N-1)S^2. \tag{18}$$

Therefore,

$$\frac{1}{N(N-1)} \sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_\ell - \mathbf{W}(\mathcal{K})]\right] \tag{19}$$

$$= \frac{1}{N(N-1)} \left[-(N-1)S^2\right] \tag{20}$$

$$= -\frac{S^2}{N}, \tag{21}$$

thus

$$K = \sum_i^m \sum_{\neq j}^m \mathbb{E}\left[[\mathbf{w}_i - \mathbf{W}(\mathcal{K})]^T [\mathbf{w}_j - \mathbf{W}(\mathcal{K})]\right] \tag{22}$$

$$= m(m-1)\frac{1}{N(N-1)} \sum_k^N \sum_{\neq \ell}^N \left[[\mathbf{w}_k - \mathbf{W}(\mathcal{K})]^T [(\mathbf{w}_l - \mathbf{W}(\mathcal{K})]\right] \tag{23}$$

$$= -m(m-1)\frac{S^2}{N}, \tag{24}$$

and substitute the value of K, the variance of $\mathbf{w}_{rand}$ is

$$\mathbb{V}\left(\mathbf{w}_{rand}\right) = \frac{N-1}{Nm}S^2 - \frac{1}{n^2}m(m-1)\frac{S^2}{N} \tag{25}$$

$$= \frac{N-m}{Nm}S^2. \tag{26}$$

If N is infinite (large enough), we can get

$$\mathbb{V}\left(\mathbf{w}_{rand}\right) = \frac{N-m}{Nm}S^2 \tag{27}$$

$$= (\frac{1}{m} - \frac{1}{N})S^2 \cong \frac{S^2}{m}. \tag{28}$$

Derive the Variance of Plain Clustering Selection. As prior work constructed, clustering selection is always applied under plain proportional allocation, where the number of sampled clients m_h from the h^{th} cluster is proportional to its cluster size N_h, i.e., $m_h = m\frac{N_h}{N}$. And we have

$$\begin{aligned}\mathbb{V}(\mathbf{w}_{cluster}) = &\sum_{h=1}^{\mathrm{H}} {Q_h}^2 \mathbb{V}\left(\mathbf{w}_h\right) \\ &+ \sum_{h(\neq j)=1}^{\mathrm{H}} \sum_{j=1}^{m_h} Q_h Q_j \operatorname{Cov}\left(\mathbf{w}_h, \mathbf{w}_j\right).\end{aligned} \tag{29}$$

For the former we have

$$\mathbb{V}\left(\mathbf{w}_h\right) = \frac{N_h - m_h}{N_h m_h}{S_h}^2, \tag{30}$$

and for the latter (covariance) we have

$$\operatorname{Cov}\left(\mathbf{w}_h, \mathbf{w}_j\right) = 0, h \neq j, \tag{31}$$

where

$${S_h}^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h} \left\|\mathbf{w}_{h_j} - \mathbf{W}_h\right\|_2^2, \tag{32}$$

thus

$$\mathbb{V}(\mathbf{w}_{cluster}) = \sum_{h=1}^{\mathrm{H}} \left(\frac{N_h - m_h}{N_h m_h}\right){Q_h}^2{S_h}^2. \tag{33}$$

Therefore we can get

$$\mathbb{V}(\mathbf{w}_{cluster}) = \sum_{h=1}^{\mathrm{H}} \left(\frac{N_h - \frac{m}{N}N_h}{N_h \frac{m}{N}N_h}\right)\left(\frac{N_h}{N}\right)^2 {S_h}^2 \tag{34}$$

$$= \frac{N-m}{Nm}\sum_{h=1}^{\mathrm{H}} \frac{N_h {S_h}^2}{N} \tag{35}$$

$$= \frac{N-m}{Nm}\sum_{h=1}^{\mathrm{H}} Q_h {S_h}^2 \tag{36}$$

$$\cong \frac{\sum_{h=1}^{\mathrm{H}} N_h {S_h}^2}{mN}. \tag{37}$$

Derive the Variance of Clustering Selection with Sample Size Re-allocation. We apply clustering selection under sample size re-allocation, where the number of sampled clients m_h from the h^{th} cluster is proportional to both cluster's size N_h and the variability of cluster measured by S_h, i.e.,

$$m_h = \frac{N_h S_h}{\sum_{h=1}^{\mathrm{H}} N_h S_h} \cdot m. \tag{38}$$

We can get

$$\begin{aligned}
\mathbb{V}\left(\mathbf{w}_{cludiv}\right) &= \sum_{h=1}^{\mathrm{H}} \left(\frac{1}{m_h} - \frac{1}{N_h}\right) {Q_h}^2 {S_h}^2 \\
&= \sum_{h=1}^{\mathrm{H}} \frac{{Q_h}^2 {S_h}^2}{m_h} - \sum_{h=1}^{\mathrm{H}} \frac{{Q_h}^2 {S_h}^2}{N_h} \\
&= \sum_{h=1}^{\mathrm{H}} \left[{Q_h}^2 {S_h}^2 \left(\frac{\sum_{h=1}^{\mathrm{H}} N_h S_h}{m N_h S_h} \right) \right] - \sum_{h=1}^{\mathrm{H}} \frac{{Q_h}^2 {S_h}^2}{N_h} \\
&= \sum_{h=1}^{\mathrm{H}} \left[\frac{1}{m} \cdot \frac{N_h S_h}{N^2} \left(\sum_{h=1}^{\mathrm{H}} N_h S_h \right) \right] - \sum_{h=1}^{\mathrm{H}} \frac{{Q_h}^2 {S_h}^2}{N_h} \\
&= \frac{1}{m} \left(\sum_{h=1}^{\mathrm{H}} \frac{N_h S_h}{N} \right)^2 - \sum_{h=1}^{\mathrm{H}} \frac{{Q_h}^2 {S_h}^2}{N_h} \\
&= \frac{1}{m} \left(\sum_{h=1}^{\mathrm{H}} Q_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{\mathrm{H}} Q_h {S_h}^2 \\
&= \frac{1}{N^2} \frac{\left(\sum_{h=1}^{\mathrm{H}} N_h S_h \right)^2}{m} - \frac{1}{N^2} \sum_{h=1}^{\mathrm{H}} N_h {S_h}^2 \\
&\cong \frac{1}{m N^2} \left(\sum_{h=1}^{\mathrm{H}} N_h S_h \right)^2 .
\end{aligned} \tag{39–46}$$

Based on all the above, We have these equations below when approximations are used,

$$\begin{aligned}
\mathbb{V}\left(\mathbf{w}_{rand}\right) &= \frac{N-m}{Nm} S^2 \qquad (47) \\
&\cong \frac{S^2}{m}, \qquad (48) \\
\mathbb{V}\left(\mathbf{w}_{cluster}\right) &= \frac{N-m}{N} \cdot \frac{\sum_{h=1}^{\mathrm{H}} N_h {S_h}^2}{mN} \qquad (49) \\
&\cong \frac{\sum_{h=1}^{\mathrm{H}} N_h {S_h}^2}{mN}, \qquad (50) \\
\mathbb{V}\left(\mathbf{w}_{cludiv}\right) &= \frac{1}{N^2} \cdot \frac{\left(\sum_{h=1}^{\mathrm{H}} N_h S_h \right)^2}{m} - \frac{1}{N^2} \sum_{h=1}^{\mathrm{H}} N_h {S_h}^2 \qquad (51) \\
&\cong \frac{1}{m N^2} \left(\sum_{h=1}^{\mathrm{H}} N_h S_h \right)^2 . \qquad (52)
\end{aligned}$$

Relationship. In order to compare $\mathbb{V}(w_{rand})$ and $\mathbb{V}(w_{cluster})$, we first attempt to express S^2 as a function of ${S_h}^2$.

$$
\begin{aligned}
(N-1)\,S^2 &= \sum_{h=1}^{\mathrm{H}} \sum_{i=1}^{m_h} \left\|\mathbf{w}_{h_i} - \mathbf{W}(\mathcal{K})\right\|_2^2 && (53)\\
&= \sum_{h=1}^{\mathrm{H}} \sum_{i=1}^{m_h} \left\|\mathbf{w}_{h_i} - \mathbf{W}_h\right\|_2^2 && (54)\\
&+ \sum_{h=1}^{\mathrm{H}} N_h \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2 && (55)\\
&= \sum_{h=1}^{\mathrm{H}} (N_h - 1)\,{S_h}^2 + \sum_{h=1}^{\mathrm{H}} N_h \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2 && (56)
\end{aligned}
$$

$$
\frac{N-1}{N} S^2 = \sum_{h=1}^{\mathrm{H}} \frac{N_h - 1}{N} {S_h}^2 + \sum_{h=1}^{\mathrm{H}} \frac{N_h}{N} \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2 \tag{57}
$$

We assume that N_h is large enough to permit the approximation for simplification

$$
\frac{N_h - 1}{N_h} \approx 1 \text{ and } \frac{N-1}{N} \approx 1 \tag{58}
$$

Thus

$$
S^2 = \sum_{h=1}^{\mathrm{H}} \frac{N_h}{N} {S_h}^2 + \sum_{h=1}^{\mathrm{H}} \frac{N_h}{N} \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2 \tag{59}
$$

Therefore

$$
\begin{aligned}
\mathbb{V}\left(\mathbf{w}_{rand}\right) = \frac{S^2}{m} &= \frac{\sum_{h=1}^{\mathrm{H}} N_h {S_h}^2}{mN} && (60)\\
&+ \frac{\sum_{h=1}^{\mathrm{H}} N_h \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2}{mN} && (61)\\
&= \mathbb{V}\left(\mathbf{w}_{cluster}\right) + \frac{\sum_{h=1}^{\mathrm{H}} N_h \left\|\mathbf{W}_h - \mathbf{W}(\mathcal{K})\right\|_2^2}{mN} && (62)
\end{aligned}
$$

which shows that

$$
\mathbb{V}\left(\mathbf{w}_{cluster}\right) \leq \mathbb{V}\left(\mathbf{w}_{rand}\right) \tag{63}
$$

Unless $\mathbf{W}_h = \mathbf{W}(\mathcal{K})$ for every h, we must have $\mathbb{V}\left(\mathbf{w}_{cludiv}\right) \leq \mathbb{V}\left(\mathbf{w}_{cluster}\right)$.
The difference is

$$
\mathbb{V}\left(\mathbf{w}_{cluster}\right) = \mathbb{V}\left(\mathbf{w}_{cludiv}\right) + \frac{1}{mN} \sum_{h=1}^{\mathrm{H}} N_h \left(S_h - \overline{S}\right)^2 . \tag{64}
$$

This shows that

$$
\mathbb{V}\left(\mathbf{w}_{cludiv}\right) \leq \mathbb{V}\left(\mathbf{w}_{cluster}\right), \tag{65}
$$

unless $S_h = \overline{S}$ for every h, i.e., the clusters have equal variability. Therefore, we get

$$
\mathbb{V}\left(\mathbf{w}_{cludiv}\right) \leq \mathbb{V}\left(\mathbf{w}_{cluster}\right) \leq \mathbb{V}\left(\mathbf{w}_{rand}\right) \tag{66}
$$

In this paper, we proposed to apply the importance selection based on the norm of the gradient to each cluster instead of random selection. Here we present the variance-reduction relationship between random selection and importance selection. Please note that this part is directly adapted from prior importance sampling work Katharopoulos and Fleuret [2018], which is not our contribution.

$$
\begin{aligned}
&\operatorname{Tr}\left(\mathbb{V}_{rand}\left[G_i\right]\right) - \operatorname{Tr}\left(\mathbb{V}_{import}\left[c_i G_i\right]\right) && (67)\\
&= \mathbb{E}_{rand}\left[\left\|G_i\right\|_2^2\right] - \mathbb{E}_{import}\left[c_i^2 \left\|G_i\right\|_2^2\right] && (68)
\end{aligned}
$$

Using the fact that $c_i = \frac{1}{N_h G_i}$, $I_i = \frac{\|G_i\|_2}{\sum_{i=1}^{N_h}\|G_i\|_2}$, $u = \frac{1}{N_h}$, we have

$$\mathbb{E}_{import}\left[c_i^2\left\|G_i\right\|_2^2\right] = \left(\frac{1}{N_h}\sum_{i=1}^{N_h}\|G_i\|_2\right)^2 \tag{69}$$

Then simplify it, we can get

$$\begin{aligned}
&\operatorname{Tr}\left(\mathbb{V}_{rand}\left[G_i\right]\right) - \operatorname{Tr}\left(\mathbb{V}_{import}\left[w_i G_i\right]\right) && (70)\\
&= \frac{1}{N_h}\sum_{i=1}^{N_h}\|G_i\|_2^2 - \left(\frac{1}{N_h}\sum_{i=1}^{N_h}\|G_i\|_2\right)^2 && (71)\\
&= \frac{\left(\sum_{i=1}^{N_h}\|G_i\|_2\right)^2}{N_h^3}\sum_{i=1}^{N_h}\left(N_h^2\frac{\|G_i\|_2^2}{\left(\sum_{i=1}^{N_h}\|G_i\|_2\right)^2} - 1\right) && (72)\\
&= \frac{\left(\sum_{i=1}^{N_h}\|G_i\|_2\right)^2}{N_h}\sum_{i=1}^{N_h}\left(I_i^2 - u^2\right) && (73)
\end{aligned}$$

Using the fact that$\sum_{i=1}^{N_h} u = 1$, we can complete the derivation.

$$\begin{aligned}
&\operatorname{Tr}\left(\mathbb{V}_{rand}\left[G_i\right]\right) - \operatorname{Tr}\left(\mathbb{V}_{import}\left[w_i G_i\right]\right) && (74)\\
&= \frac{\left(\sum_{i=1}^{N_h}\|G_i\|_2\right)^2}{N_h}\sum_{i=1}^{N_h}\left(I_i - u\right)^2 && (75)\\
&= \left(\frac{1}{N_h}\sum_{i=1}\|G_i\|_2\right)^2 N_h\|I - u\|_2^2 && (76)
\end{aligned}$$

□

3 ADDITIONAL EXPERIMENTS

3.1 INFLUENCE OF THE SAMPLING RATIO

Table 2: Final test accuracy of multiple FL algorithms with different sampling schemes under convex model on MNIST, FMNIST, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 50$, $\eta = 0.05$, $B = 50$.

	Methods	MNIST		FMNIST	
		IID	non-IID	IID	non-IID
q=0.1	Random	86.8 ±0.0	79.3 ±0.6	75.9 ±0.0	62.6 ±1.0
	SCAFFOLD	84.2 ±0.2	78.8 ±1.3	73.5 ±0.0	70.7 ±0.0
	Importance	90.9 ±0.0	87.2 ±1.2	82.5 ±0.1	73.2 ±1.6
	Cluster	90.91 ±0.0	88.0 ±0.4	82.6 ±0.1	74.3 ±2.0
	HCSFed	90.9 ±0.0	**89.0 ±0.0**	82.5 ±0.1	**78.8 ±0.0**
q=0.2	Random	88.4 ±0.0	83.2 ±0.3	78.6 ±0.1	71.6 ±1.1
	SCAFFOLD	85.2 ±0.0	86.6 ±0.2	74.2 ±0.0	71.0 ±0.0
	Importance	90.8 ±0.0	87.5 ±1.0	82.6 ±0.1	75.1 ±2.0
	Cluster	90.89 ±0.0	88.6 ±0.3	82.5 ±0.1	77.3 ±1.1
	HCSFed	91.0 ±0.0	**89.1 ±0.1**	**82.5 ±0.1**	**78.9 ±0.1**
q=0.3	Random	89.3 ±0.0	86.1 ±0.2	80.1 ±0.1	71.3 ±0.2
	SCAFFOLD	84.9 ±0.1	87.0 ±0.1	74.4 ±0.0	71.1 ±0.0
	Importance	90.8 ±0.0	88.4 ±0.4	82.6 ±0.0	75.4 ±1.9
	Cluster	90.85 ±0.0	89.1 ±0.1	82.5 ±0.0	77.4 ±0.5
	HCSFed	90.9 ±0.0	**89.0 ±0.0**	82.6 ±0.1	**78.9 ±0.0**
q=0.5	Random	90.0 ±0.0	87.2 ±0.1	81.2 ±0.0	75.5 ±0.3
	SCAFFOLD	85.0 ±0.1	87.2 ±0.1	74.5 ±0.0	70.9 ±0.0
	Importance	90.9 ±0.0	89.0 ±0.2	82.5 ±0.0	77.2 ±1.0
	Cluster	90.87 ±0.0	89.0 ±0.1	82.5 ±0.0	78.7 ±0.4
	HCSFed	90.9 ±0.0	**89.0 ±0.0**	82.5 ±0.0	**78.9 ±0.0**

Table 3: Final test accuracy of multiple FL algorithms with different sampling schemes under non-convex model on MNIST, FMNIST and CIFAR-10, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 50$ for MNIST and FMNIST, $nSGD = 80$ for CIFAR-10, $\eta = 0.05$, $B = 50$.

	Methods	MNIST		FMNIST		CIFAR-10		
		IID	non-IID	IID	non-IID	IID	$\alpha = 0.01$	$\alpha = 0.001$
q=0.1	Random	87.0 ±0.0	59.7 ±1.4	87.6 ±0.1	76.9 ±0.2	40.3 ±0.2	25.8 ±0.4	20.5 ±0.5
	Importance	92.9 ±0.1	71.2 ±9.5	90.4 ±0.1	85.1 ±2.2	66.0 ±0.2	39.5 ±3.3	24.9 ±2.6
	Cluster	92.9 ±0.0	73.6 ±3.7	90.6 ±0.2	88.9 ±4.6	65.5 ±0.3	37.2 ±3.9	30.0 ±4.7
	HCSFed	92.9 ±0.0	**83.3 ±0.0**	90.5 ±0.1	**92.0 ±0.2**	65.7 ±0.3	**41.2 ±1.8**	**38.8 ±0.6**
q=0.2	Random	89.3 ±0.0	70.8 ±1.6	88.8 ±0.0	80.5 ±0.5	49.8 ±0.4	29.7 ±0.3	25.1 ±0.2
	Importance	92.9 ±0.0	72.9 ±9.0	90.3 ±0.0	90.8 ±0.6	65.7 ±0.2	40.7 ±2.9	30.5 ±2.0
	Cluster	92.9 ±0.0	80.3 ±1.7	90.4 ±0.1	90.7 ±1.0	65.6 ±0.3	42.0 ±1.4	33.0 ±3.6
	HCSFed	92.8 ±0.0	**83.5 ±0.1**	90.4 ±0.1	**92.1 ±0.2**	65.4 ±0.3	**41.6 ±0.9**	**39.2 ±2.0**
q=0.3	Random	90.1 ±0.0	73.4 ±2.1	89.4 ±0.0	83.7 ±0.3	54.8 ±0.4	34.9 ±0.6	26.6 ±0.5
	Importance	92.9 ±0.0	76.3 ±3.5	90.6 ±0.1	90.5 ±1.3	65.7 ±0.2	42.0 ±2.3	32.6 ±3.4
	Cluster	92.9 ±0.0	81.8 ±1.5	90.5 ±0.1	91.6 ±0.5	66.1 ±0.3	43.0 ±1.0	34.6 ±2.1
	HCSFed	92.8 ±0.0	**83.3 ±0.1**	90.2 ±0.1	**92.2 ±0.1**	65.4 ±0.3	**42.3 ±0.6**	**39.7 ±0.7**
q=0.5	Random	91.4 ±0.0	80.9 ±1.3	90.1 ±0.1	87.9 ±0.3	61.1 ±0.2	39.4 ±0.4	30.5 ±0.6
	Importance	92.9 ±0.0	80.8 ±3.8	90.6 ±0.1	90.5 ±1.5	65.8 ±0.3	43.2 ±1.2	35.4 ±1.7
	Cluster	92.9 ±0.0	83.5 ±0.2	90.4 ±0.1	92.0 ±0.2	65.9 ±0.4	44.0 ±0.6	36.3 ±1.0
	HCSFed	92.9 ±0.0	**83.3 ±0.1**	90.4 ±0.1	**92.0 ±0.2**	65.7 ±0.5	**42.4 ±0.7**	**39.8 ±0.6**

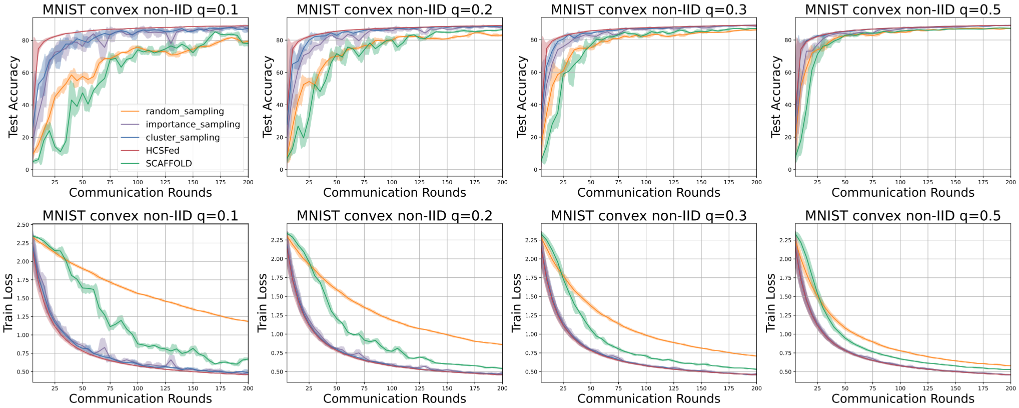


Figure 1: Impact of sampling ratio q on the performance with convex model. We compare HCSFed with simple random sampling, importance sampling, cluster sampling, SCAFFOLD on MNIST under non-IID, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$.

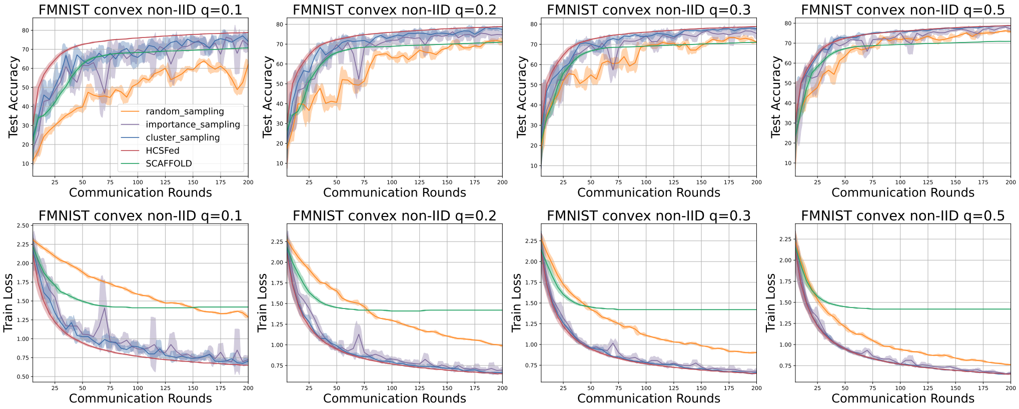


Figure 2: Impact of sampling ratio q on the performance with convex model. We compare HCSFed with simple random sampling, importance sampling, cluster sampling, SCAFFOLD on FMNIST under non-IID, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$.

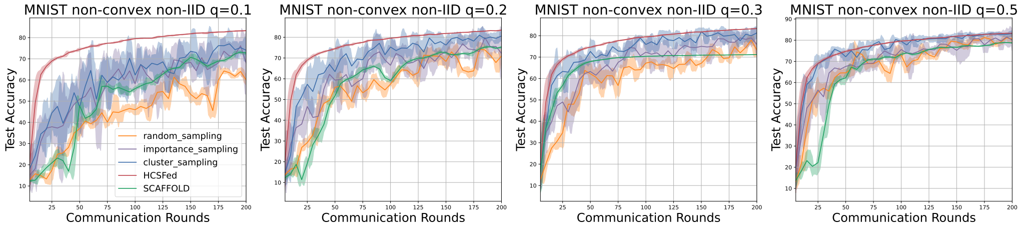


Figure 3: Impact of sampling ratio q on the performance with non-convex model. We compare HCSFed with simple random sampling, importance sampling, cluster sampling, SCAFFOLD on MNIST under non-IID, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$.

3.2 EXTRA EXPERIMENTS ON FEDNOVA

We carry out extra experiments on FedNova, a modified FL algorithm, to further verify the compatibility of our sampling scheme. We use all datasets mentioned above and take different distributions into consideration. As illustrated in Figure 6, our sampling scheme achieves superb performance on FedNova, especially under heterogeneity.

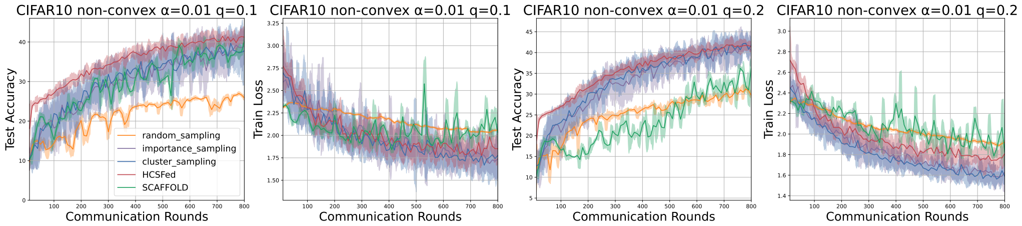


Figure 4: Impact of sampling ratio q on the performance with non-convex model. We compare `HCSFed` with simple random sampling, importance sampling, cluster sampling, SCAFFOLD on CIFAR-10, using a Dirichlet Distribution with $\alpha = 0.01$, setting parameters $q \in \{0.1, 0.2\}$, $N = 100$, $nSGD = 80$, $\eta = 0.05$, $B = 50$.

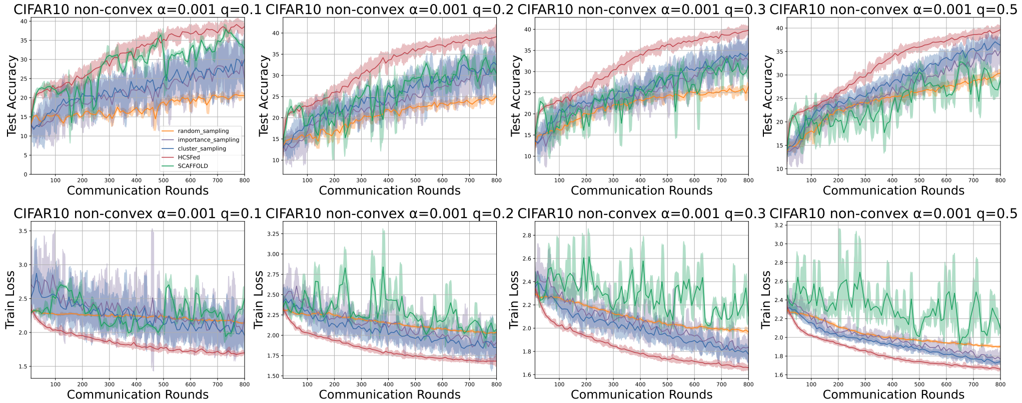


Figure 5: Impact of the heterogeneity on the performance with non-convex model. We compare `HCSFed` with simple random sampling, importance sampling, cluster sampling and SCAFFOLD on CIFAR-10, using a Dirichlet Distribution with $\alpha = 0.001$, setting parameters $q \in \{0.1, 0.2, 0.3, 0.5\}$, $N = 100$, $nSGD = 80$, $\eta = 0.05$, $B = 50$.

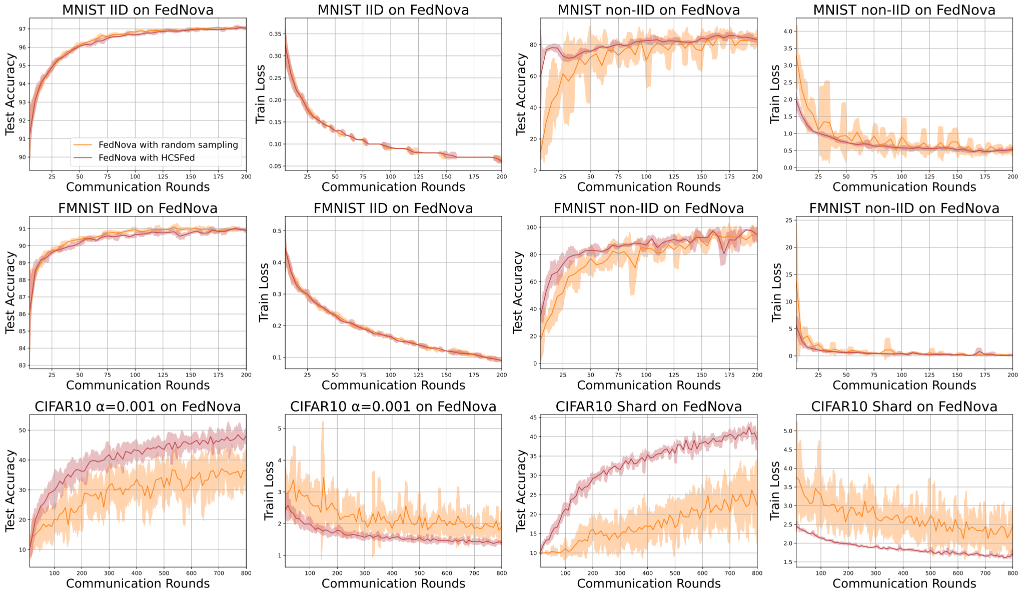


Figure 6: Results on the MNIST, FMNIST and CIFAR10 under FedNova. We compare `HCSFed` with simple random sampling, using a Dirichlet Distribution with $\alpha = 0.001$ and Shard, setting parameters $q = 0.1$, $N = 100$, $nSGD = 50$ for MNIST and FMNIST, $nSGD = 80$ for CIFAR-10, $\eta = 0.05$, $B = 50$.

4 WIDELY USED DATASETS IN PREVIOUS INFLUENTIAL FL OPTIMIZATION WORKS

As for the datasets, we would like to state that constructing the heterogeneous dataset to verify our ideas with a non-iid setting deserves more attention in FL optimization. Meanwhile, we would like to make a fair comparison with the previous Influential FL optimization works shown in table 4, so we choose the widely used dataset including MNIST, FMNIST, and CIFAR-10.

Table 4: Comparison of widely used datasets in previous influential FL optimization works.

Articles	Datasets they use
Cluster samplingFraboni et al. [2021]	MNIST, CIFAR-10
Importance sampingChen et al. [2022]	FEMNIST, Shakespeare
SCAFFOLDKarimireddy et al. [2020]	EMNIST
FedNovaWang et al. [2020]	Synthetic Federated dataset, CIFAR-10
FedProxLi et al. [2020]	MNIST, FEMNIST, Shakespeare, Sent140

References

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=8GvRCWKHIL`.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 2020.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, pages 3321–3363, 2013.